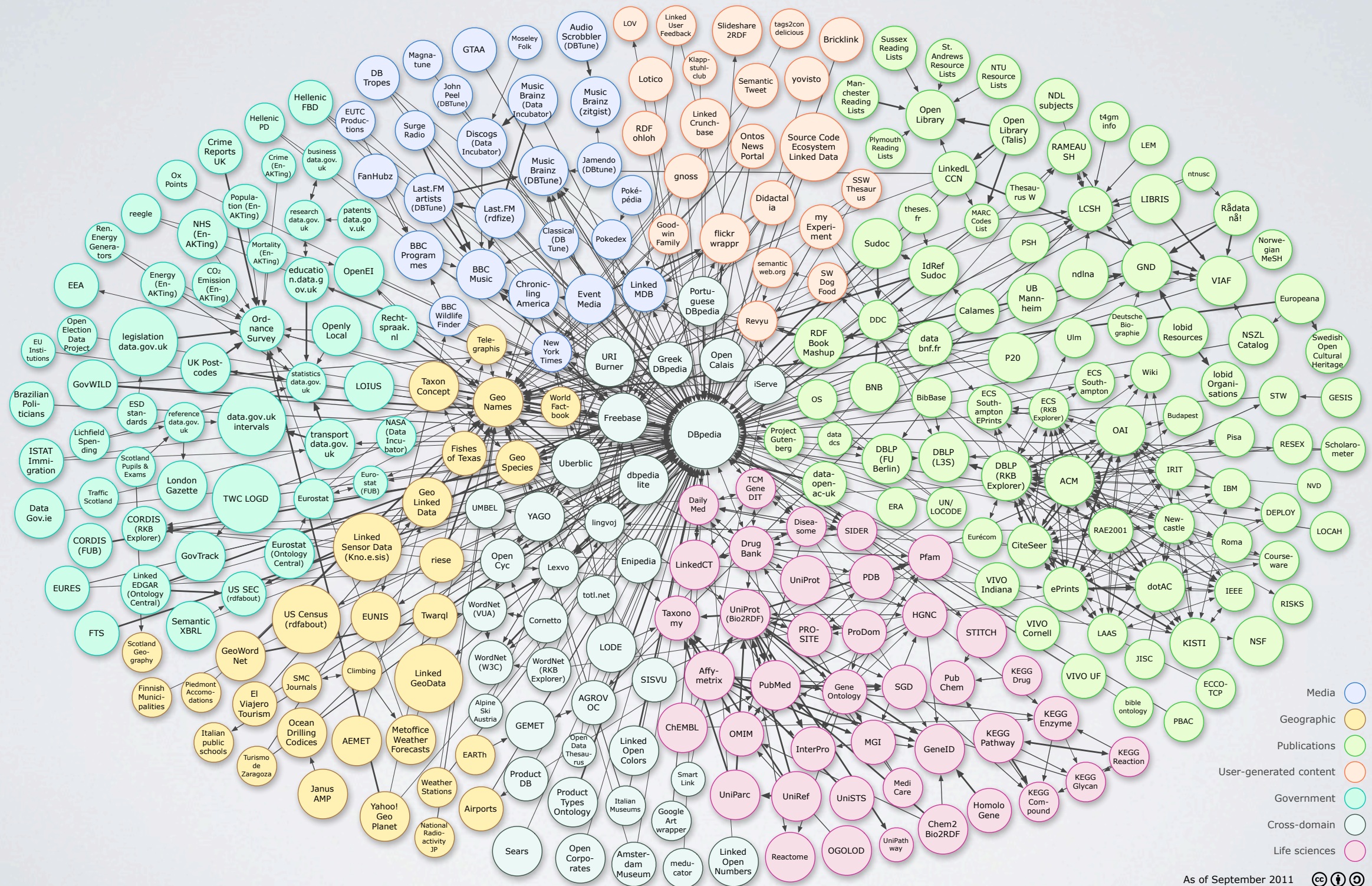


NEUROWIKI:

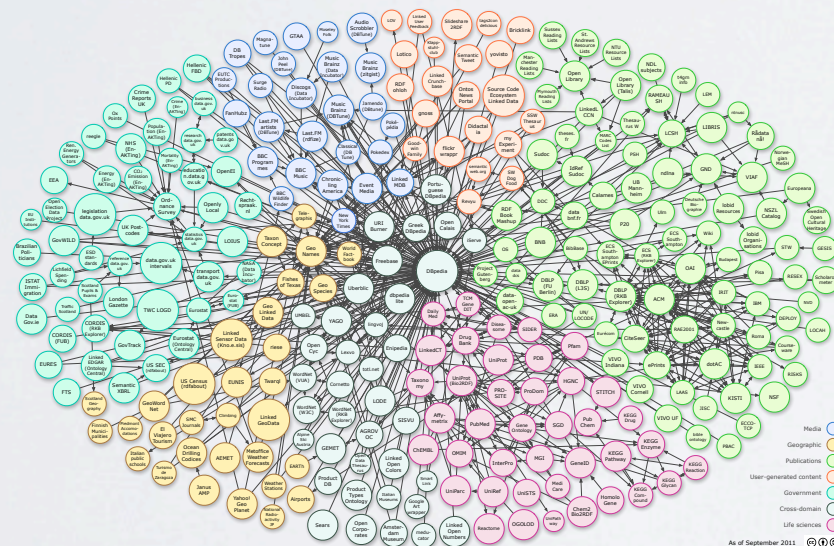
HOW WE INTEGRATED LARGE DATASETS INTO SMW WITH LDIF

OBLIGATORY SLIDE



LINKED DATA CHALLENGES

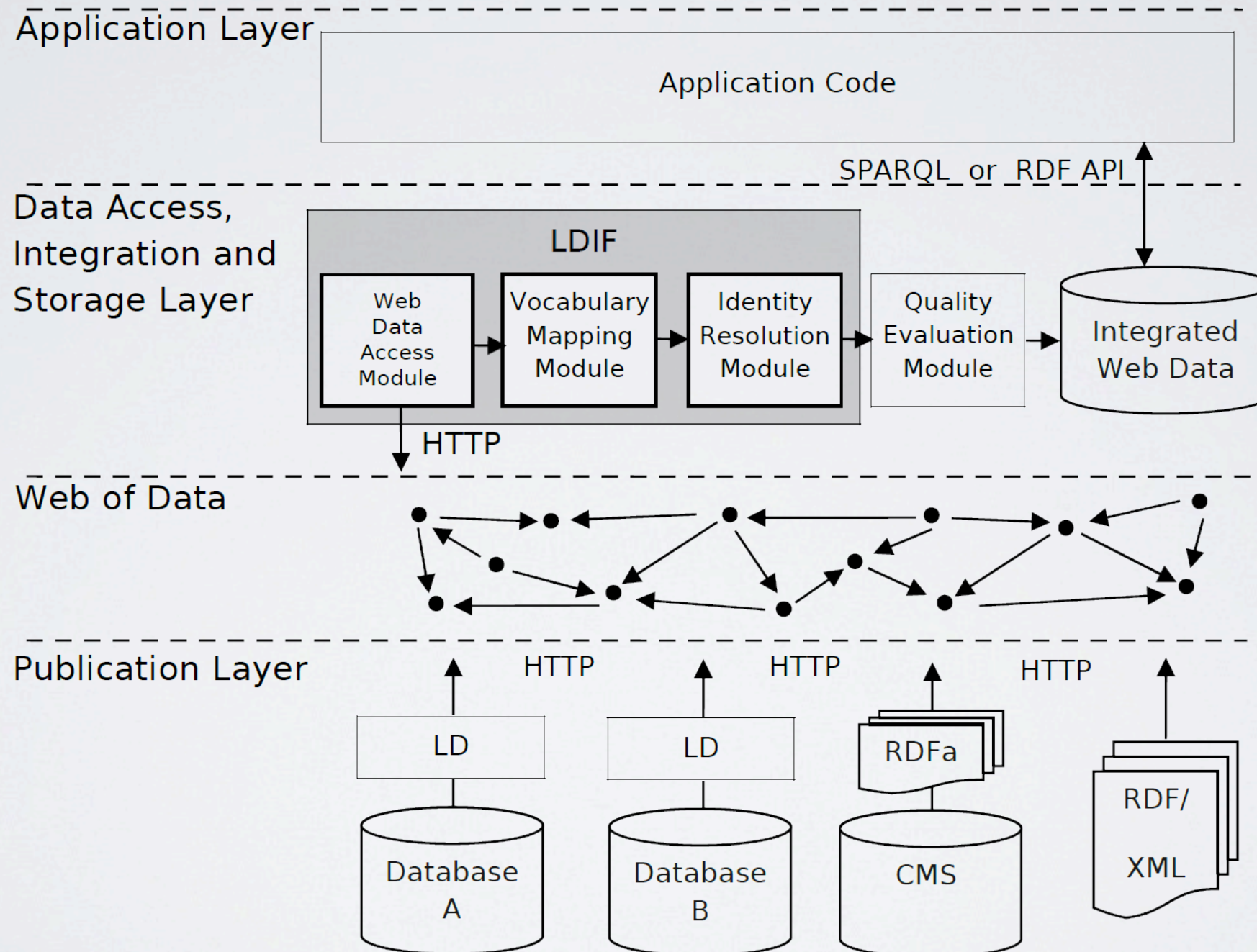
- obtaining the data and keeping it up to date
- data sources use a wide range of different RDF vocabularies to represent data about the same type of entity
- different URIs identify the same real-world entity



LDIF – LINKED DATA INTEGRATION FRAMEWORK

- manages data download and update
- translates heterogeneous data into a single local target vocabulary
- replaces URI aliases with a single target URI on the client side
- outputs the results to files or a quad store with provenance
- LDIF is available standalone and as part of Ontoprise TSC in conjunction with the SMW+ Linked Data Extension
- Supported in part by Vulcan Inc. as part of its Project Halo and by the EU FP7 project LOD2 - Creating Knowledge out of Interlinked Data (Grant No. 257943).

LINKED DATA APPLICATION ARCHITECTURE



CURRENT LDIF GOAL

Scalability!

	25M	50M	100M
Load and build entites for R2R	128.1 <i>sec</i>	297.2 <i>sec</i>	1059.7 <i>sec</i>
R2R data translation	169.9 <i>sec</i>	515.0 <i>sec</i>	1109.2 <i>sec</i>
Build entities for Silk	15.3 <i>sec</i>	36.8 <i>sec</i>	107.4 <i>sec</i>
Silk Identity Resolution	103.0 <i>sec</i>	568.5 <i>sec</i>	2954.9 <i>sec</i>
Final URI rewriting	8.1 <i>sec</i>	27.0 <i>sec</i>	65.0 <i>sec</i>
Overall execution time	7.0 <i>min</i>	24.0 <i>min</i>	88.3 <i>min</i>

R2R INTRODUCTION

A Problem-Solution Approach

PROBLEM: NAME MISMATCH

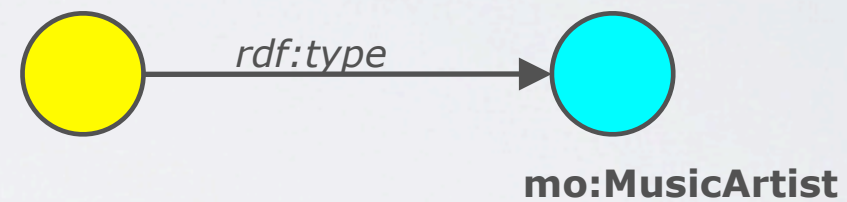
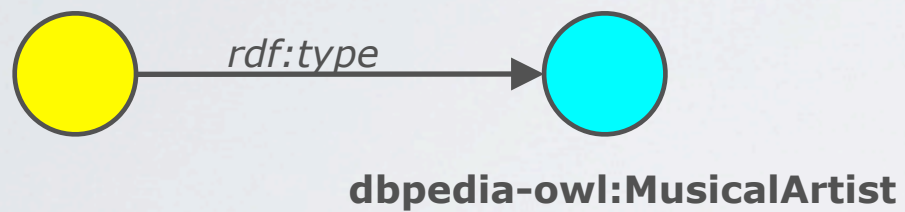
Example: Music Artist

- Music Ontology: `mo:MusicArtist`
- DBpedia ontology: `dbpedia-owl:MusicalArtist`

SOLUTION TO NAME MISMATCH

SourcePattern

TargetPattern



PROBLEM: STRUCTURE MISMATCH

- Example: DBpedia vs. Factbook

- DBpedia ontology:

- Property: `dbpedia-owl:leaderName`

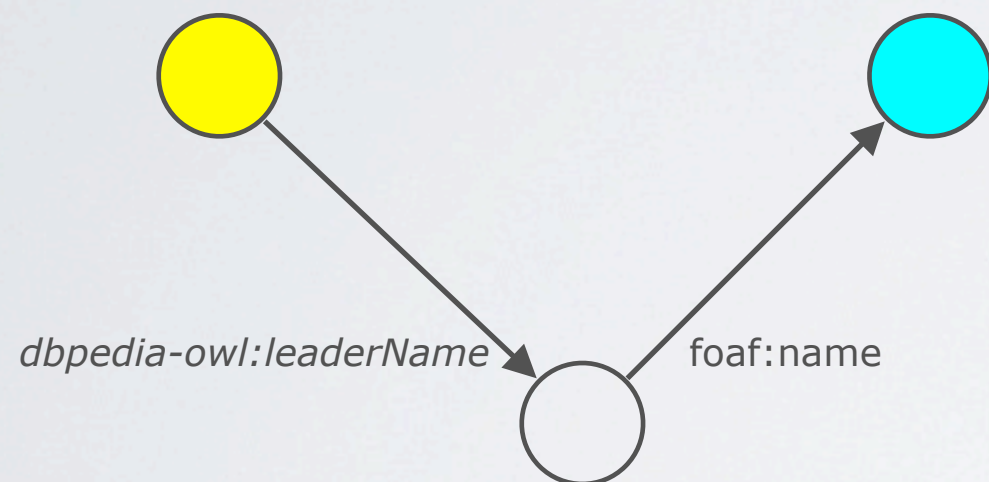
- Property: `foaf:name`

- Factbook Ontology:

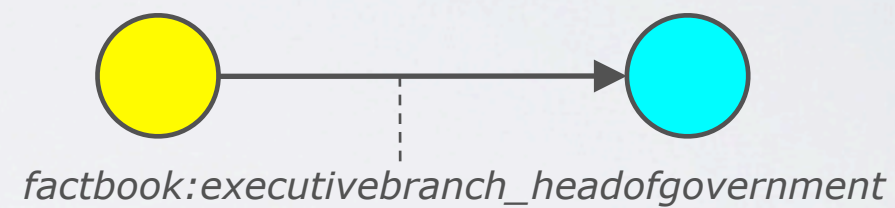
- Property: `factbook:executivebranch_headofgovernment`

SOLUTION TO STRUCTURE MISMATCH

SourcePattern



TargetPattern



MODIFIERS

- Datatype modifier: “120” => “120”^^xsd:double
- Language modifier
 - Used to specify a language tag for a literal
- URI modifier
 - To convert a value into a URI
 - Example: "http://dbpedia.org" to <http://dbpedia.org>
- Literal modifier
 - To convert a URI into a literal

PROBLEM: SCHEMA MISMATCH

- Example: Music Ontology vs. DBpedia

- Music Ontology:

- Class: `mo:Record`

- Property: `mo:release_type`

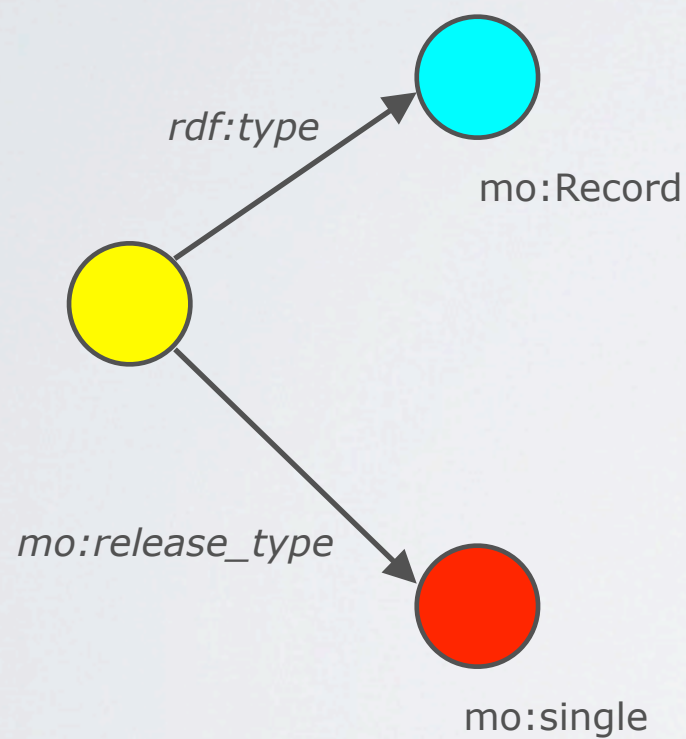
- Instance: `mo:single`

- DBpedia ontology:

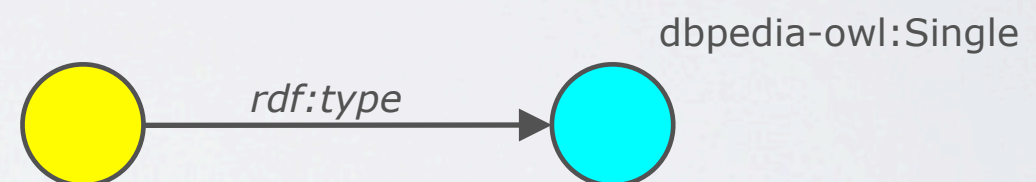
- Class: `dbpedia-owl:Single`

SOLUTION: RESTRICTIONS

SourcePattern



TargetPattern



PROBLEM:VALUE MISMATCH

- Example: PharmGKB vs. SMW

- PharmGKB Ontology:

Property: pharmgkb:Drugbank_Id

Value example: <http://chem2biordf.org/.../DB00317>

- SMW ontology:

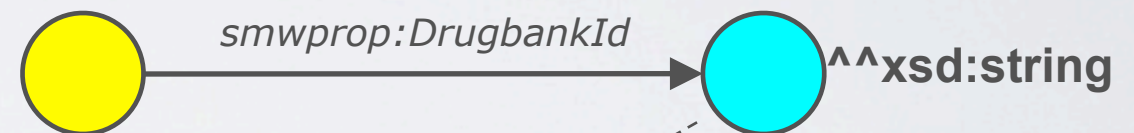
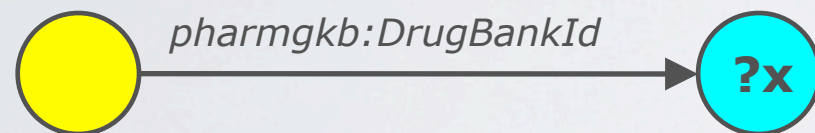
Property: smwprop:DrugBankId

Value example: "DB00317"^^xsd:string

SOLUTION: TRANSFORMATIONS

SourcePattern

TargetPattern



`regexToList('http://chem2bio2rdf.org/drugbank/resource/drugbank_drug/(.+?)', ?x)`

SILK INTRODUCTION

By Example

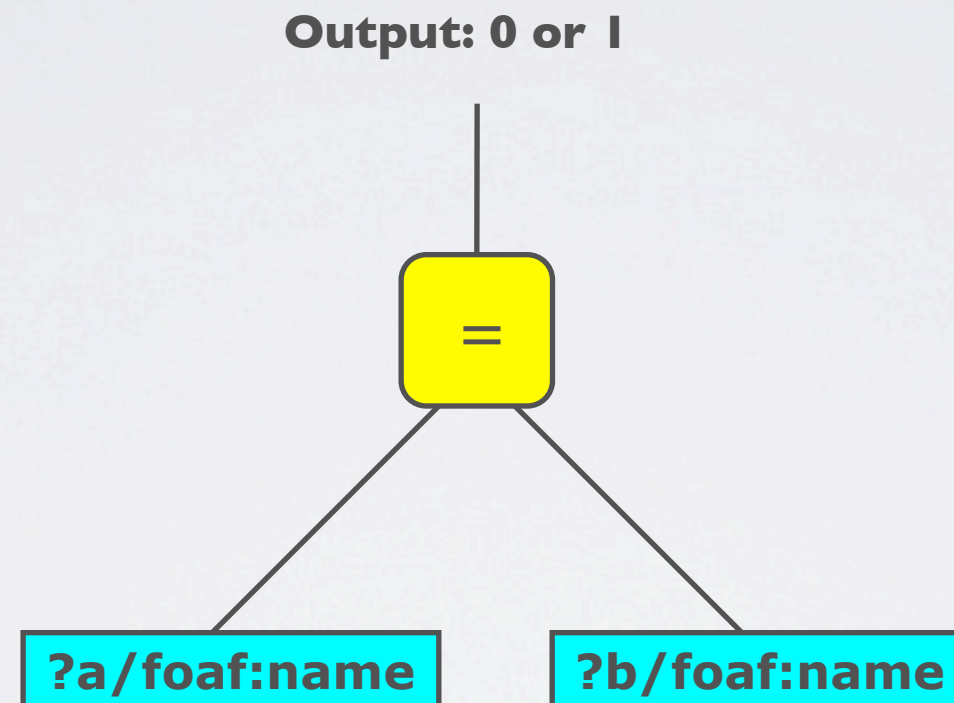
SILK: LINKAGE RULES

- Specify when two entities should be considered the same
- Elements of a linkage rule:
 - Restrict the set of entities you want to compare
 - Pick the relevant values
 - Compare the values
 - Aggregate results of different comparators
 - Results above a defined threshold are considered as matches

HOWTO CHOOSE AND COMPARE VALUES

- Example use case: Mainstream music bands
 - Band names are usually trademarked
 - Assumption: band names are unique

SOLUTION: PATH INPUT AND COMPARATORS



Entities represented by ?a and ?b are restricted to mo:MusicGroup

EXAMPLES OF OTHER COMPARATORS

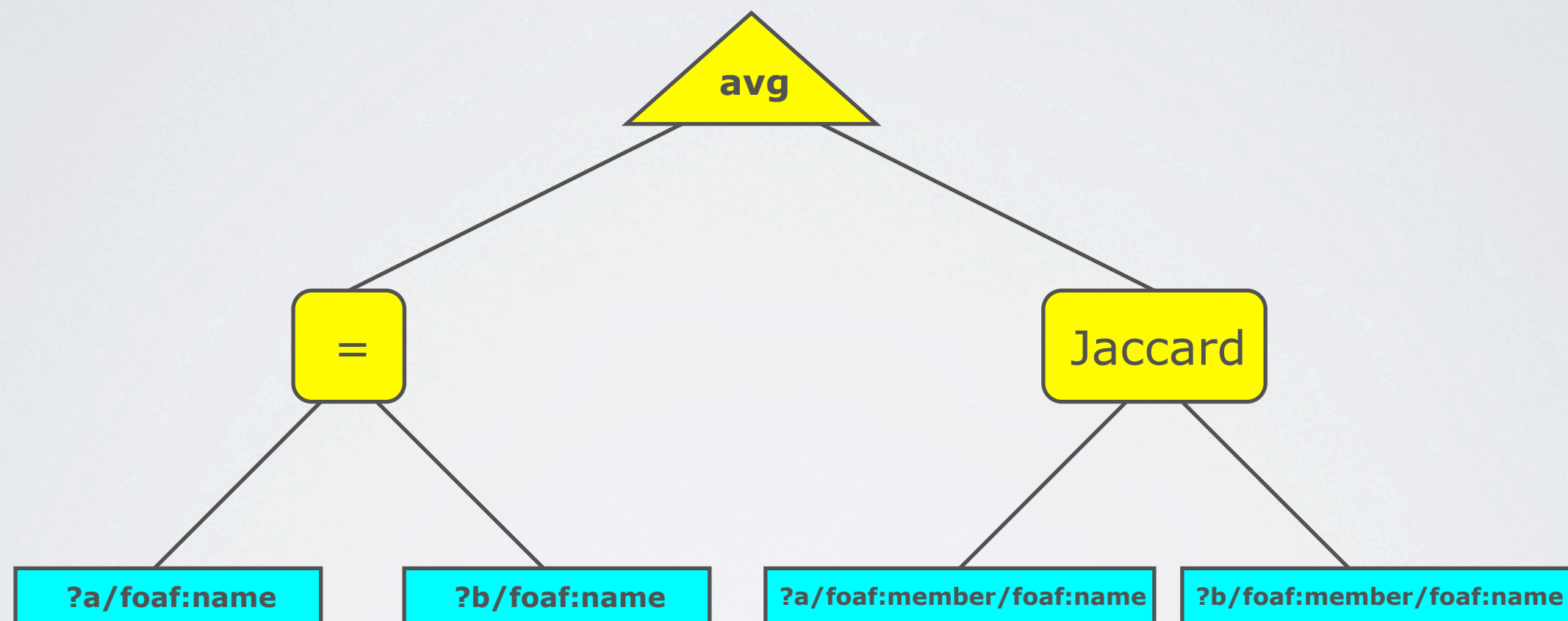
- String based similarities like Edit Distance
- Token based similarities like Jaccard's Coefficient
- Data type specific: geo-coordinates, date types etc.

PROBLEM: NON-UNIQUE NAMES

- Example: Local music bands
 - Band names are usually not trademarked
 - Assumption: Band names are NOT unique
- Not enough to only compare band names!

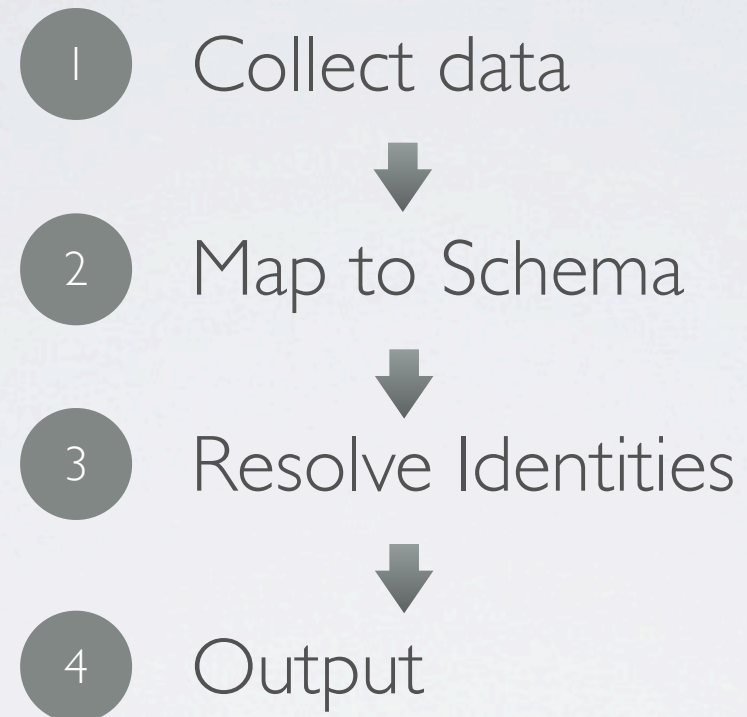
SOLUTION: COMBINE COMPARATORS WITH AGGREGATION

Output: Between 0 and 1



Entities represented by ?a and ?b are restricted to mo:MusicGroup

LDIF PIPELINE



- Parallelized on one machine
- Next release will add Hadoop support

LDIF PIPELINE



Supported data sources:

- RDF dumps (various formats)
- SPARQL Endpoints
- Crawling Linked Data



Name	Data source	Last import	Change frequency	Imported	Status message		
SiderMapped	http://mes.smw-lde-eu.s3.amazonaws.com/sider_dump_fixed.nt.bz2	9/15/11 8:51 AM	-	yes		(Re-)Import	Update
KEGGGeneMapped	http://mes.smw-lde-eu.s3.amazonaws.com/kegg_genes_20101018_100.nt.bz2	9/15/11 9:13 AM	-	yes		(Re-)Import	Update
SiderOriginal	http://mes.smw-lde-eu.s3.amazonaws.com/sider_dump_fixed.nt.bz2	9/15/11 8:50 AM	-	yes	Unable to find a matching description that maps http://www.example.org/smw-lde/smwDatasources/SiderOriginal to wiki	(Re-)Import	Update
KEGGGeneOriginal	http://mes.smw-lde-eu.s3.amazonaws.com/kegg_genes_20101018_100.nt.bz2	9/15/11 9:13 AM	-	yes	Unable to find a matching description that maps http://www.example.org/smw-lde/smwDatasources/KEGGGeneOriginal to wiki	(Re-)Import	Update
DrugbankOriginal	http://mes.smw-lde-eu.s3.amazonaws.com/drugbank_dump_fixed.nt.bz2	9/15/11 6:29 AM	-	yes	Unable to find a matching description that maps http://www.example.org/smw-lde	(Re-)Import	Update

Data source definition for DiseasomeOriginal

The Linked Data source definition was parsed successfully. The following values will be stored:

ID	DiseasomeOriginal
Label	DiseasomeOriginal
Data dump location	< http://mes.smw-lde-eu.s3.amazonaws.com/diseasome_dump_fixed.nt.bz2 >

Data source definition for DiseasomeMapped

The Linked Data source definition was parsed successfully. The following values will be stored:

ID	DiseasomeMapped
Label	DiseasomeMapped
Data dump location	< http://mes.smw-lde-eu.s3.amazonaws.com/diseasome_dump_fixed.nt.bz2 >

Data source definition for DBpediaToDiseasome

The Linked Data source definition was parsed successfully. The following values will be stored:

ID	DBpediaToDiseasome
Label	DBpediaToDiseasome
Data dump location	< http://mes.smw-lde-eu.s3.amazonaws.com/dbpedia_diseasome.nt.bz2 >

Drugbank

Note: drugbank_dump_fixed.nt has the following two lines were commented out from http://www4.wiwiss.fu-berlin.de/drugbank/drugbank_dump.nt.bz2 :

- <<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00013> > <<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/rxlistLink> >
<<http://www.rxlist.com/cgi/pharmclips2.cgi?keyword=%20Abbokinase%AE> > .



[Show RichTextEditor]



== Diseasome ==

```
{{#sourcedefinition:
  id = DiseasomeOriginal
  Label = DiseasomeOriginal
  DataDumpLocation = <http://mes.smw-lde-eu.s3.amazonaws.com/diseasome_dump_fixed.nt.bz2>
}}
```

```
{{#sourcedefinition:
  id = DiseasomeMapped
  Label = DiseasomeMapped
  DataDumpLocation = <http://mes.smw-lde-eu.s3.amazonaws.com/diseasome_dump_fixed.nt.bz2>
}}
```

```
{{#sourcedefinition:
  id = DBpediaToDiseasome
  Label = DBpediaToDiseasome
  DataDumpLocation = <http://mes.smw-lde-eu.s3.amazonaws.com/dbpedia_diseasome.nt.bz2>
}}
```

== Drugbank ==

Note: drugbank_dump_fixed.nt has the following two lines were commented out from [http://www4.wiwiiss.fu-berlin.de/drugbank/drugbank_dump.nt.bz2 http://www4.wiwiiss.fu-berlin.de/drugbank/drugbank_dump.nt.bz2]:

Press Ctrl+Alt+Space to use auto-completion. (Ctrl+Space in IE)

Please note that all contributions to Neuro Wiki may be edited, altered, or removed by other contributors. If you do not want your writing to be edited mercilessly, then do not submit it here.

You are also promising us that you wrote this yourself, or copied it from a public domain or similar free resource (see [Neuro Wiki:Copyrights](#) for details).

Do not submit copyrighted work without permission!

LDIF PIPELINE

- 1 Collect data
- ↓
- 2 Map to Schema
- ↓
- 3 Resolve Identities
- ↓
- 4 Output

Using R2R

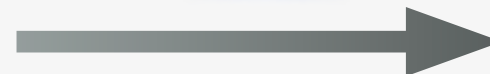
KEGG:gene

KEGG:hasPathway




wiki:Gene

wiki:IsInvolvedIn














[My dashboard](#) ▾

[Articles and data](#) ▾

 [LODMappings](#)
[Special page](#)
[More](#) ▾

All R2R Mappings

ID	From	To	Edit
DiseasomeMapped_to_Wiki_Mapping_1	DiseasomeMapped	Wiki	
DrugbankMapped_to_Wiki_Mapping_1	DrugbankMapped	Wiki	
DrugbankMapped_to_Wiki_Mapping_2	DrugbankMapped	Wiki	
KEGGGeneMapped_to_Wiki_Mapping_1	KEGGGeneMapped	Wiki	
KEGGGeneMapped_to_Wiki_Mapping_4	KEGGGeneMapped	Wiki	
KEGGPathwayMapped_to_Wiki_Mapping_1	KEGGPathwayMapped	Wiki	
KEGGPathwayMapped_to_Wiki_Mapping_5	KEGGPathwayMapped	Wiki	
SiderMapped_to_Wiki_Mapping_1	SiderMapped	Wiki	
SiderMapped_to_Wiki_Mapping_3	SiderMapped	Wiki	

 [New R2R Mapping](#)



DiseasomeMapped_to_Wiki_Mapping_1

	Name	Source	Target	Edit
▼	mp:Disease	diseasome:diseases	smwcat:Disease	
	mp:associatedGene	diseasome:associatedGene	smwprop:associatedGene	
	mp:diseaseLabel	rdfs:label	rdfs:label	
	mp:possibleDrug	diseasome:possibleDrug	smwprop:possibleDrug	
	New Property Mapping			
▼	mp:Gene	diseasome:genes	smwcat:Gene	
	mp:geneLabel	rdfs:label	rdfs:label	
	New Property Mapping			
	New Class Mapping			

Back to Overview

Remove Mapping



Edit Property Mapping

Mapping Tree

Mapping Source



Prefix Definitions



Source Pattern



Target Pattern



Transformation

mp:diseaseLabel



?SUBJ rdfs:label ?label



?SUBJ rdfs:label ?labelTransformed



?labelTransformed = regexToList('(.+?)(?:, [0-9]+)?', ?label)

Save

Remove

Cancel



Back to Overview

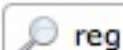
Remove Mapping



Edit Property Mapping

Transformation

?labelTransformed = regexToList('(.+?)(?:, [0-9]+)?', ?label)



reg

String functions

split(regex, stringarg)

regexToList(regex, stringarg)

replaceAll(thisRegex, withThatString, inThisString)

Arithmetic functions

List functions

XPath functions

xpath_matches(s, pattern)

xpath_replace(s, pattern, replacement)

regexToList(regex, stringarg)

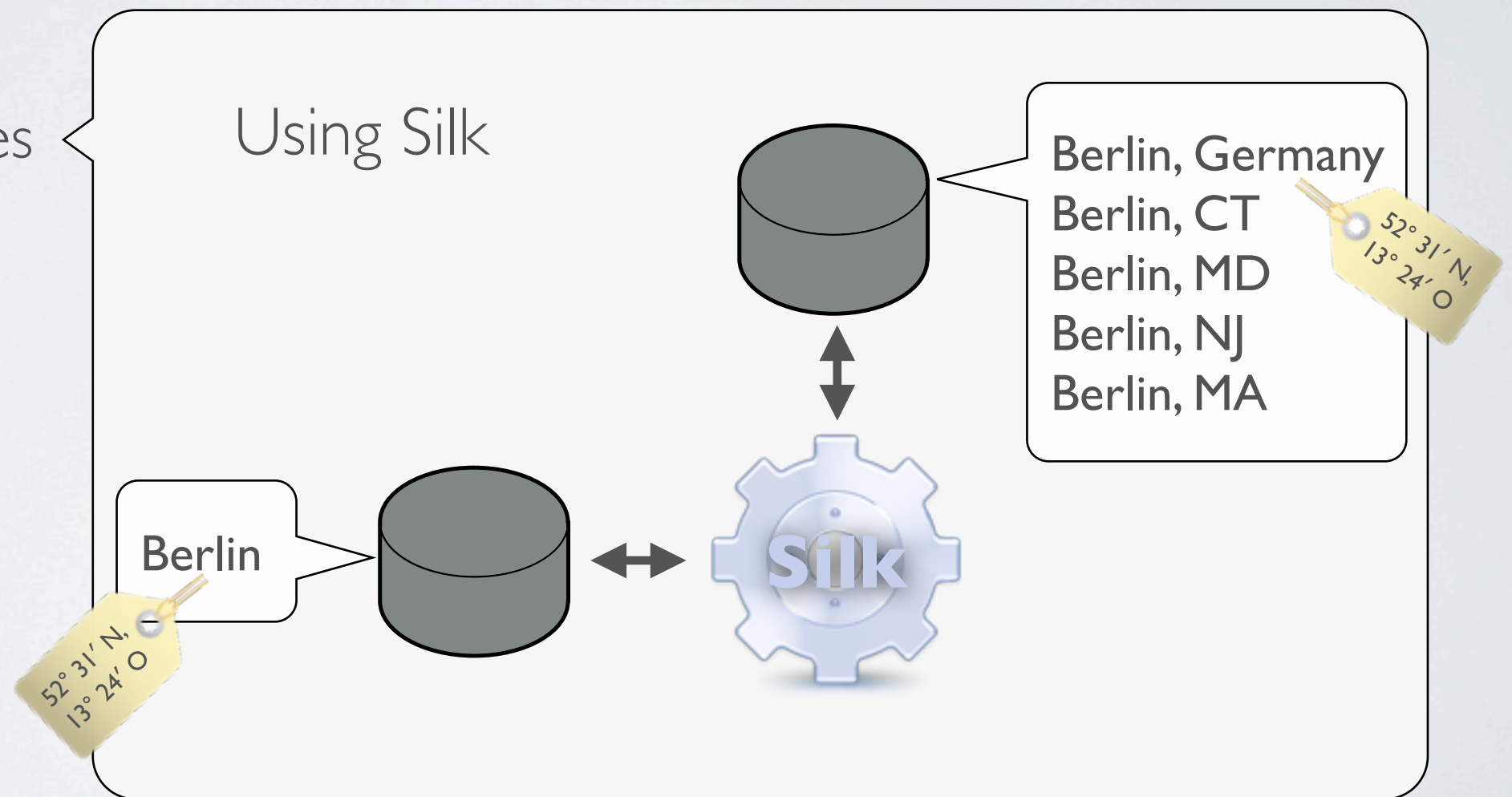
Returns a list of strings as specified by the regex

Save

Cancel

LDIF PIPELINE















































- 1 Collect data
- ↓
- 2 Map to Schema
- ↓
- 3 Resolve Identities
- ↓
- 4 Output



Silk Workbench

Workspace: KEGG_Pathway_to_Wiki_Silk_mapping

About

-   Project
-  Disease_to_Wiki_Silk_mapping  Prefixes  Task  Output  Link Spec  Remove
-  Drugbank_to_Wiki_Silk_mapping  Prefixes  Task  Output  Link Spec  Remove
-  KEGG_GENES_to_Wiki_Silk_mapping  Prefixes  Task  Output  Link Spec  Remove
-  KEGG_Pathway_to_Wiki_Silk_mapping  Prefixes  Task  Output  Link Spec  Remove
-  SOURCE  Edit
-  diseases  Metadata  Open  Remove
-  genes  Metadata  Open  Remove
-  pathways  Metadata  Open  Remove
-  SIDER_to_Wiki_Silk_mapping  Prefixes  Task  Output  Link Spec  Remove

Silk Workbench

Workspace: KEGG_Pathway_to_Wiki_Silk_mapping

Editor: genes

Generate Links

Reference Links

About

Export as Silk-LS

Help

Property Paths

Source: SOURCE

Restriction: ?b rdf:type smwcat:Gene .

(custom path)

?b/smwprop:UniprotId

?b/smwprop:EntrezGeneId

?b/smwprop:MgiMarkerAccessionId

?b/smwprop:name

Target: TARGET

Restriction: ?a rdf:type smwcat:Gene .

(custom path)

?a/smwprop:UniprotId

?a/smwprop:EntrezGeneId

?a/smwprop:MgiMarkerAccessionId

?a/smwprop:name

Transformations

Alpha reduce

Concatenate

Convert Charset

Logarithm

Lower case

Comparators

Date

DateTime

Dice coefficient

Equality

Geographical distance

Aggregators

Average

Euclidian distance

Geometric mean

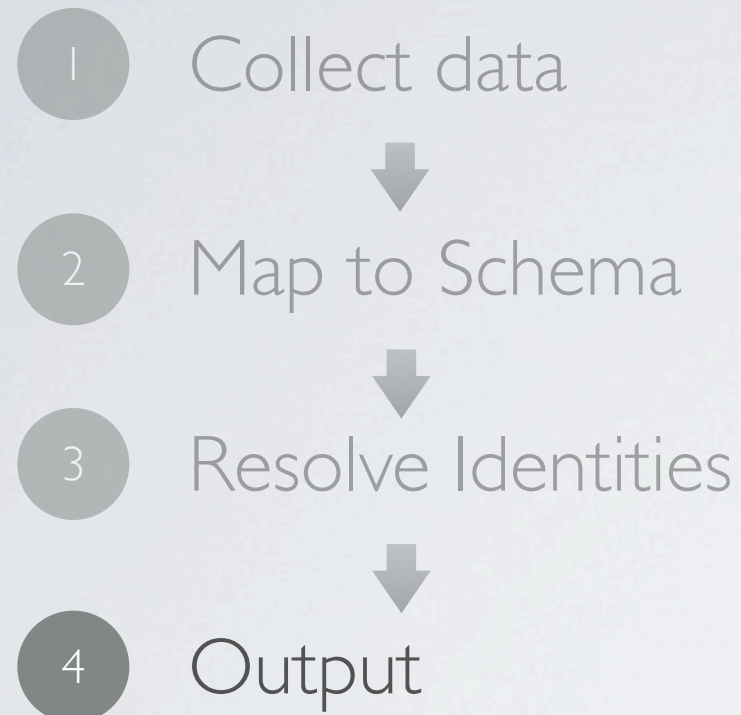
Maximum

Minimum



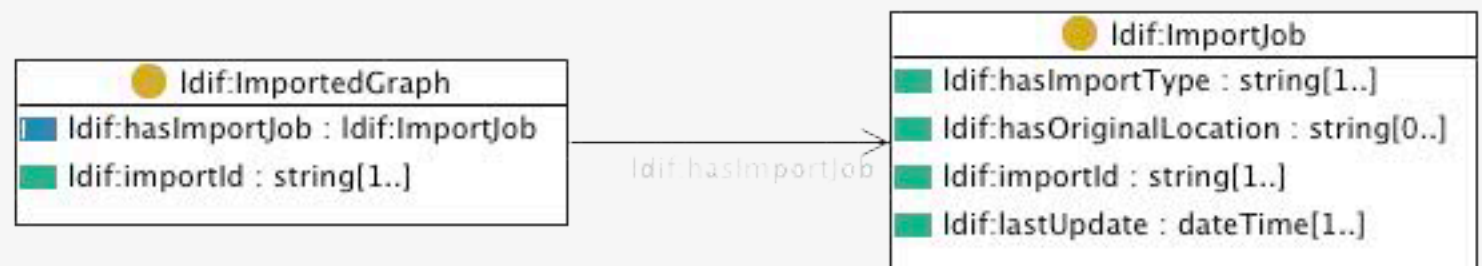
Link Limit: unlimited 

LDIF PIPELINE



Output options:

- N-Quads
- N-Triples
- SPARQL Update Stream
- Includes provenance



Q & A

THANKS!

- We're looking for first adopters!
- Website: <http://bit.ly/ldifweb>
- Google Group: <http://bit.ly/ldifgroup>
- SMW+ Linked Data Extension: <http://bit.ly/ldifsmw>